The Estimation of the Econometric Model of Milk Yield per Cow: A Support Vector Machine Regression Approach

Põldaru R. a and Roots J. b

a Estonian Agricultural University, 51014 Tartu, Estonia, preet@eua.ee
b Estonian Agricultural University, 51014 Tartu, Estonia, jroots@eua.ee

Abstract

In this paper we give an overview of the basic ideas underlying Support Vector (SV) machines for regression and function estimation. A summary of currently used algorithms for training SV machines will be presented. Application of SVM regression for estimating parameters of econometric model of milk yield per cow in Estonian farms will be considered and possibilities of application of SVM regression in rural areas are discussed. Studies on implementation of SVM regression methods (algorithms) and software packages in agricultural research and business must be extended.

Key words: data mining, support vector machines regression, econometric models, milk yield per cow

1 Introduction

The amount of information that is being created and stored in the computer is increasing very rapidly. Everyday transactions, protocols, and documents are being stored in the computer and automated monitoring systems create vast information repositories.

With the advent of the Internet, these information resources have become available to individuals and companies regardless of national borders and constraints of time and space. As a consequence, information overload is rapidly becoming the new plague of the information society. It is, therefore, becoming increasingly important to provide effective tools to help users organize, manage, understand, and access large repositories of information. New data analysis procedures provided by current data mining (DM) have substantially changed the situation in the field of data processing (DP). The situation in data mining is the most challenging.

Data mining, often called knowledge discovery in databases (KDD), is the process of discovery of useful information from large collections of data. It has common frontiers with several fields including Data Base Management (DBM), Artificial Intelligence (AI), Machine Learning (ML), Pattern Recognition (PR), and Data Visualisation (DV).

The researchers of the Institute of Informatics of the EAU have investigated the possibilities of using some new DM methods and also have some experience in implementing algorithms used in DM packages (Bayesian statistical methods, neural networks, principal components method, decision trees and rules (e.g. CART – classification and regression trees), association rules discovery, fuzzy regression methods and support vector machine regression). The results of the research are published in many papers and conference theses (see references in Põldaru, Roots, and Ruus (2003)).

Current paper provides an overview about the support vector machines regression (SVMR), describes the potential implementation of it in rural areas and discusses the implementation of this method for analysing the dairy sector in Estonia (of econometric model of milk yield per cow). The data are an
unbalanced panel of milk producers drawn from the FADN (Farm Accountancy Data Network) database of Estonian milk producers. The parameters are estimated on the basis of alternative models of SVM regression using linear and nonlinear model. The results are compared mutually and with results of ordinary linear regression. For linear and non-linear model (parameter) estimation the SVM module of Programming Environment R was used. R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

2 Methods of investigation

Support Vector Machines (SVMs) have been successfully applied to a number of applications ranging from particle identification, face identification and text categorisation to engine knock detection, bioinformatics and database marketing. The approach is systematic and properly motivated by statistical learning theory (Vapnik 1998). Training (model parameter estimation) involves optimisation of a convex cost function: there are no false local minimum to complicate the learning process. The approach has many other benefits, for example, the model constructed has an explicit dependence on the most informative patterns in the data (the support vectors), hence interpretation is straightforward and data cleaning could be implemented to improve performance. SVMs are the most well known of a class of algorithms which use the idea of kernel substitution and which we will broadly refer to as kernel methods.

Suppose we are given statistical data \{(x_1, y_1), \ldots, (x_n, y_n)\}. These might be, for instance, exchange rates for some currency measured at subsequent days together with corresponding econometric indicators. Our goal in SVM regression is to find a function \(f(x)\) that has at most \(\varepsilon\) deviation from the actually obtained targets \(y_i\) for all the (training) data, and at the same time, is as flat as possible. SVM regression uses the \(\varepsilon\) - insensitive loss function. If the deviation between the actual and predicted value is less than \(\varepsilon\), then the regression function is not considered to be in error.

Thus mathematically we would like \(-\varepsilon \leq a \cdot x_i + b - y_i \leq \varepsilon\). Geometrically, we can visualize this as a band or tube of size \(2\varepsilon\) around the hypothesis function \(f(x)\) and any points outside this tube can be viewed as errors (see Fig. 1). All training (data) points \((x_i, y_i)\) for which \(|f(x_i) - y_i| \geq \varepsilon\) are known as support vectors; it is only these points that determine the parameters of \(f(x)\). In other words, we do not care about errors as long as they are less than \(\varepsilon\) but will not accept any deviation larger than this. We begin by describing the case of simple linear functions \(f(x)\), taking the form

\[
f(x) = a \cdot x + b
\]

Flatness in the case of (1) means that one seeks small \(a\). One way to ensure this is to minimize the Euclidean norm, i.e. \(a^2\). Formally we can write this problem as a convex optimization problem by requiring:

minimise \(\|e\|^2\)

subject to \[\begin{align*}
y_j - a \cdot x_i - b &\leq \varepsilon \\
-a \cdot x_i + b - y_j &\leq \varepsilon
\end{align*}\]

The tacit assumption in (2) is (was), that the convex optimization problem is feasible.

Analogously to the `soft margin` loss function in (Vapnik, 1998), one can introduce slack variables \(\xi_i, \xi_i^+\) to cope with otherwise infeasible constraints of the optimization problem (2). Hence we arrive at the formulation stated in (Vapnik, 1998):

minimise \(\|e\|^2 + C \cdot \sum_{i=1}^{n} (\xi_i + \xi_i^+)\)
subject to

\[
\begin{align*}
y_j - a \cdot x_i - b & \leq \varepsilon + \xi_i \\
a \cdot x_i + b - y_i & \leq \varepsilon + \xi_i^* \\
|\xi_i|, |\xi_i^*| & \geq 0
\end{align*}
\] (3)

Fig. 1. Plot of \( f(x) = a \cdot x + b \) versus \( x \) with \( \varepsilon \)-insensitive tube. Points outside tube are errors.

The constant \( C > 0 \) determines the trade off between the flatness of \( f(x) \) and the amount up to which deviations larger than \( \varepsilon \) are tolerated. The formulation above corresponds to dealing with a so called \( \varepsilon \)-insensitive loss function \( |\xi|_\varepsilon \) described by

\[
|\xi|_\varepsilon = \begin{cases} 
0 \text{...if } |\xi| \leq \varepsilon \\
|\xi| - \varepsilon \text{...otherwise} 
\end{cases}
\] (4).

Fig. 1 depicts the situation graphically.

It turns out that the optimization problem (3) can be solved more easily in its dual formulation. Moreover, as we will see next, the dual formulation provides the key for extending SV machine to nonlinear functions. Hence we will use a standard dualization method utilizing Lagrange multipliers.

In the case of the Lagrangian dual (supporting) optimization problem we need to optimise:

maximise

\[
\left\{-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i - \alpha_i^*) \cdot (\alpha_j - \alpha_j^*) \cdot x_i \cdot x_j - \varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i \cdot (\alpha_i - \alpha_i^*)\right\}
\]

subject to

\[
\begin{align*}
\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) &= 0 \\
0 &\leq \alpha_i, \alpha_i^* \leq C
\end{align*}
\] (5)

where \( \alpha_i \) and \( \alpha_i^* \) are Lagrangian multiplier.

The value of regression parameter \( a \) and predicted value \( f(x) \) can be calculated as follows

\[
a = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot x_i
\] (6)
and  

\[ f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot x_i \cdot x + b \]  

(7).

This is the so-called Support Vector expansion, i.e. \(a\) can be completely described as a linear combination of the training patterns \(x_i\). In a sense, the complexity of a function's representation by SVs is independent of the dimensionality of the input space \(X\), and depends only on the number of SVs. Moreover, the complete algorithm can be described in terms of dot products between the data. Even when evaluating \(f(x)\) we need not compute \(a\) explicitly (although this may be computationally more efficient in the linear setting). These observations will come handy for the formulation of a nonlinear extension.

A final note has to be made regarding the sparsity of the SV expansion. Only for \(|f(x_i) - y_i| \geq \varepsilon\) the Lagrange multipliers may be nonzero, or in other words, for all samples inside the \(\varepsilon\)-tube in Fig. 1 the \(\alpha_i\) and \(\alpha_i^*\) vanish: for \(|f(x_i) - y_i| \leq \varepsilon\). Therefore we have a sparse expansion of \(a\) in terms of \(x_i\) (i.e. we do not need all \(x_i\) to describe \(a\)). The examples that come with non-vanishing coefficients are called Support Vectors.

The next step is to make the SV algorithm nonlinear. This, for instance, could be achieved by simply preprocessing the training patterns \(x_i\) by a map into some feature space \(F\), as described in (Vapnik 1998) and then applying the standard SV regression algorithm.

We must first define a mapping from the space \(X\) of regressors to the possibly infinite dimensional hypothesis space \(H\), in which an inner product \(<,>\) is defined. We formally describe this map as

\[ \Phi : X \rightarrow H \quad \text{or} \quad x \mapsto \Phi(x) \]  

(8).

We choose to limit our choice of regression function \(f(x)\) to the class of functions which can be expressed as inner products in \(H\), taken between some weight vector \(a\) and the mapped regressor \(\Phi(x)\):

\[ f(x) = \langle a, \Phi(x) \rangle + b \]  

(9).

The regression function in the hypothesis space is consequently linear, and thus the non linear regression problem of estimating \(f(x)\) has become a linear regression problem in the hypothesis space \(H\). Note that the mapping \(\Phi(\cdot)\) need never be computed explicitly; instead, we use the fact that if \(H\) is the reproducing kernel Hilbert space induced by \(k(\cdot, \cdot)\), then writing \(\Phi(x) = k(x, \cdot)\), we get

\[ \langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j). \]  

(10)

The latter requirement is met for kernels fulfilling the Mercer conditions (Vapnik 1998). These conditions are satisfied for a wide range of kernels, including Gaussian radial basis functions (RBF)

\[ k(x_i, x_j) = \exp\left(-\gamma \cdot (x_i - x_j)^2\right) \]  

(11)

and polynomial function

\[ k(x_i, x_j) = \left(-\gamma \cdot x_i \cdot x_j + g\right)^d \]  

(12).

We emphasise that the feature space need never be defined explicitly, since only the kernel is used in SVM regression algorithms. Indeed, it is possible for multiple feature spaces to be included by a single kernel.

Consequently, this allows us to rewrite the SV algorithm (formulas (5) … (7)) as follows
maximise

\[
\left\{ -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\alpha_i^* - \alpha_j^*) \cdot (\alpha_j^* - \alpha_j^*) \cdot k(x_i, x_j) - \varepsilon \cdot \sum_{i=1}^{n} \alpha_i^* + \sum_{i=1}^{n} y_i \cdot (\alpha_i - \alpha_i^*) \right\}
\]

subject to

\[
\begin{align*}
\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) &= 0 \\
0 &\leq \alpha_i, \alpha_i^* \leq C
\end{align*}
\]

and regression parameter \( a \) and predicted value \( f(x) \) can be calculated as follows

\[
a = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot \Phi(x_i)
\]

\[
f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \cdot k(x_i, x) + b
\]

The difference to the linear case is that \( a \) is no longer explicitly given. However due to the theorem of Fischer-Riesz (see e.g. (Riesz and Nagy, 1955)) it is already uniquely defined in the weak sense by the dot products \( \langle a, \Phi(x) \rangle \). Also note that in the nonlinear setting, the optimization problem corresponds to finding the flattest function in feature space, not in input space.

3 Results and discussion

Next we consider the potential implementation of SVM regression in rural areas and discuss the implementation of this method for estimating an econometric model of the milk yield per cow data.

The econometric model of milk yield is defined by

\[
y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4 + b_5 \cdot x_5
\]

where,

- \( y \) represents average milk yield per cow (kg)
- \( x_1 \) represents the Number of Cows in sample farm;
- \( x_2 \) represents the Total Labor Input per Cow (hours);
- \( x_3 \) represents the Purchased (additional) concentrated feed per Cow (krons);
- \( x_4 \) represents the Total costs of feed per Cow (krons);
- \( x_5 \) represents the Invested Capital per hectare (krons).

The characteristics of the data are reported in Table 1.

For linear and non-linear model (parameter) estimation the SVM module (Meyer 2003) of Programming Environment R (Venables W. N., Smith D. M 2003) was used.

Table 2 presents the estimates of parameters of the linear econometric model. The parameters are estimated on the basis of alternative models of SVM regression using linear model (16).

Let us consider the variants in Table 2. The table shows that OLS estimates and SVM regression estimates are very similar. Table 2 reports also the values of coefficients of determination. For all model variants (variants 1 – 4) the values of coefficient of determination are significant and do differs little from coefficient for OLS.
Table 1. Data summary statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Y</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5078,3</td>
<td>94,9</td>
<td>236,2</td>
<td>1423,3</td>
<td>6882,7</td>
<td>16324,7</td>
</tr>
<tr>
<td>Median</td>
<td>5000,0</td>
<td>40</td>
<td>226,8</td>
<td>991,5</td>
<td>6036,5</td>
<td>9735,2</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>1072,1</td>
<td>109,9</td>
<td>98,5</td>
<td>1319,3</td>
<td>3589,7</td>
<td>18404,8</td>
</tr>
<tr>
<td>Minimum</td>
<td>2447</td>
<td>5</td>
<td>36,3</td>
<td>0</td>
<td>614,0</td>
<td>423,7</td>
</tr>
<tr>
<td>Maximum</td>
<td>8640</td>
<td>519</td>
<td>505,5</td>
<td>7194,4</td>
<td>22243,6</td>
<td>115659</td>
</tr>
</tbody>
</table>

The total number of observations \( n = 401 \).

Table 2. SVM estimates for parameters of econometric model

<table>
<thead>
<tr>
<th>Variable</th>
<th>SVM1</th>
<th>SVM2</th>
<th>SVM3</th>
<th>SVM4</th>
<th>OLS</th>
</tr>
</thead>
<tbody>
<tr>
<td>x2</td>
<td>-1,061</td>
<td>-1,413</td>
<td>-1,929</td>
<td>-1,821</td>
<td>-1,389*</td>
</tr>
<tr>
<td>x7</td>
<td>2,405</td>
<td>1,882</td>
<td>1,859</td>
<td>1,734</td>
<td>2,258*</td>
</tr>
<tr>
<td>x9</td>
<td>0,395</td>
<td>0,39</td>
<td>0,37</td>
<td>0,325</td>
<td>0,388*</td>
</tr>
<tr>
<td>x11</td>
<td>0,104</td>
<td>0,117</td>
<td>0,131</td>
<td>0,132</td>
<td>0,117*</td>
</tr>
<tr>
<td>x17</td>
<td>0,00279</td>
<td>0,00285</td>
<td>0,00211</td>
<td>0,00517</td>
<td>0,00295</td>
</tr>
<tr>
<td>R^2</td>
<td>0,543</td>
<td>0,545</td>
<td>0,540</td>
<td>0,533</td>
<td>0,547</td>
</tr>
<tr>
<td>Number of SV</td>
<td>358</td>
<td>191</td>
<td>79</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>epsilon</td>
<td>0,1</td>
<td>0,5</td>
<td>0,9</td>
<td>1,2</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>gamma</td>
<td>0,2</td>
<td>0,2</td>
<td>0,2</td>
<td>0,2</td>
<td></td>
</tr>
</tbody>
</table>

* coefficient is statistically significant

Table 2 reports also the values of numbers of support vectors and the values of epsilon (\( \varepsilon \)) and cost parameter C for various variants.

Table 2 shows that the value of coefficients of determination depends from the value of \( \varepsilon \). When the value of \( \varepsilon \) increase then the coefficients of determination decrease. The number of SV also decreases while value of \( \varepsilon \) increases.

It should be mentioned, that in the last case (variant SVM4) the number of SV equals only to 35. Consequently, the estimates of econometric model parameters are completely described by these 35 observations. The other 366 observations don’t influence the values of model parameters. At the same time the values of parameters for that variant don’t differ substantially from values of other variants.

From the considered models the most suitable for practical use (estimates are positive) is the last one model (SMV4 in Table 2).

The problem of regression models discussed so far is that they are only linear. Next we discuss the possibilities of using non-linear SVM regression models for estimating the econometric model of milk yield per cow. The main difference to the linear case is that regression coefficients \( a_i \) are no longer explicitly given and the modelling results may be used only for predicting values for regression model.

When using SVM for any given task, it is always necessary to specify a set of parameters. These parameters include such as whether you are interested in regression estimation or pattern recognition, what kernel you are using, what scaling is to be done on the data, etc. The summary of (given) parameter set for various variants of the non-linear model is reported in Table 3. Two other parameters (parameter g and d in formula (12)) have constant values for all variants: \( g = 5 \), and \( d = 3 \).

Previous studies (Põldaru et al. 2004) show that the non-linear SVM models are sensitive to “overfitting”. Most influential parameters are kernel type and parameter gamma. Our study shows that the parameter set in Table 3 give most acceptable results. Table 3 presents also result summaries of the results of various model alternatives. Summary characteristics for various alternatives are number of support vectors and coefficient of determination \( R^2 \).

Let us discuss the summary characteristics in Table 3.
Table 3. Parameter set for various alternatives of the non-linear model

<table>
<thead>
<tr>
<th>Variant</th>
<th>Specified set of parameters</th>
<th>Summary characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kernel type</td>
<td>Epsilon*</td>
</tr>
<tr>
<td>var1</td>
<td>Polynomial</td>
<td>0,1</td>
</tr>
<tr>
<td>var2</td>
<td>Polynomial</td>
<td>0,5</td>
</tr>
<tr>
<td>var3</td>
<td>Polynomial</td>
<td>0,5</td>
</tr>
<tr>
<td>var4</td>
<td>Polynomial</td>
<td>0,5</td>
</tr>
<tr>
<td>var5</td>
<td>Polynomial</td>
<td>0,5</td>
</tr>
<tr>
<td>var6</td>
<td>Radial</td>
<td>0,5</td>
</tr>
</tbody>
</table>

*parameters epsilon, gamma and C are specified for standardized data

For different alternatives the number of support vectors are different. The number of support vector depends mainly on value of parameter epsilon ($\epsilon$). The values of the coefficient of determination $R^2$ are higher than in linear model case (Table 2). The minimal value in Table 3 (0,594) is higher than maximal value in Table 2 (0,547). Consequently, the non-linear SVM regression models work well.

Next we compute and analyze the partial derivatives for cow numbers ($x_1$) and concentrated feed per cow ($x_3$) for considered alternatives. The values of partial derivatives are computed from predicted values numerically. The graph of derivatives with respect to independent variables for cow numbers is shown in Fig. 2.

![Graph of partial derivatives with respect to values of independent variables for cow numbers.](image1)

Fig. 2. Graphs of partial derivatives with respect to values of independent variables for cow numbers.

The dot line on the graphs presents OLS regression coefficient (see Table 2).

![Graph of partial derivatives with respect to values of independent variables for concentrated feed.](image2)

Fig. 3. Graphs of partial derivatives with respect to values of independent variables for concentrated feed.
Looking at the graphs we may now summarise the following conclusions:

- Some relations are essentially nonlinear. For example, the partial derivative of $y$ with respect to independent variable $x_1$ (cow numbers) for the alternative var1 has a maximum (Fig. 2). The graphs of partial derivatives for other variants are also nonlinear. Consequently, for the same variant (var1) we observe the case of “overfitting” and that variant can not be recommended for practical use.

- From the economic point of view the derivative for concentrated feed, as for production factor or resource, must be decreasing. While graphs of partial derivatives for concentrated feed (Fig. 3) for alternatives (var1 and var6) are not decreasing those alternatives are not acceptable. For alternative var6 the graph has a maximum and for alternative var1 the graph rises. But at the same time in the cases of alternative var2, var4 and var5 the graphs are decreasing and, consequently, those variants are acceptable from economic point of view.

- Comparing the graphs of derivatives with the regression coefficient (dot line on graph) we may conclude that correct (acceptable) non-linear models are more informative and more beneficial.

- The non-linear SVM regression models with lower value of epsilon and radial kernel function are more sensitive to “overfitting” (see var1 and var6).

4 Conclusions

SVM regression provides a new approach to the problem of parameter estimation of linear and especially nonlinear econometric models. In this paper we gave a brief exposition of SVM regression and their flexibility in handling economic data. Different SVM regression models are used for estimation of the econometric model of milk yield per cow in Estonian farms. The discussion may now be summarized in the following conclusions:

- Application of the SVM classification and regression in many fields of science and engineering (including econometrics) are rapidly increasing.

- The SVM regression models may be used for estimating the parameters of linear and nonlinear econometric models.

- The SVM regression estimates and the least square estimates of econometric model of milk yield per cow are similar, whereas the estimates for some independent variables are essentially equivalent.

- Using of polynomial kernel function gave more acceptable results.

- The nonlinear SVM regression models are sensitive to “overfitting”.

- The suitable parameter selection allows diminish the “overfitting” problem.

This analysis has demonstrated that interesting new methods can be implemented for parameter estimation of econometric models. We hope that this paper will encourage the use of SVM regression for econometric analysis.

References


